PRIVACY PRESERVING IN DATA STREAM USING SLIDING WINDOW METHOD

Ankit Jasoliya [1] | Tejal Patel [2]

[1] PG Student, Department of Information Technology, Parul Institute of Engineering &Technology, Vadodara, India - 390011.

[2] Assistant Professor, Department of Information Technology, Parul Institute of Engineering &Technology, Vadodara, India - 390011.

ABSTRACT

Data mining gets valuable knowledge from huge amounts of data. In latest, data streams are new type of data, which are completely different from existing static data. The property of data streams are: Data has timing preference; data distribution changes constantly with time; the amount of data is large; here data flows in and out is very quickly; and on the spot reply is necessary. Existing algorithm is designed for the static database. If the data changes, it would be compulsory to rescan the whole dataset, which takes to more computation time and providing late respond to the user. Here we studied the problem of privacy-preserving data mining and many techniques have been find. However, existing techniques for privacy-preserving data mining are designed for static databases and are not suitable for dynamic data. When need to perform computation at that time to providing privacy. So the privacy preservation problem of data streams mining is very big issue. Here we generate data stream using MOA and then provide privacy to this stream. We proposed algorithms which take over the existing process of data streams classification to achieve more privacy preservation and accuracy.

KEYWORDS: Data mining, Data Perturbation, Data Stream, MOA, Weka, Privacy, Hoeffding Tree, Classification.

1. INTRODUCTION

Data Mining is defined as getting information (meaning full data) from large sets of data. We can also say that data mining is the process of getting knowledge from large amount of data. There is a large amount of data available in the Information World. This data is of no use until it is transformed into meaning-full information. It is necessary to analyze this large amount of data and getting meaning-full information from it[12].

Information is right now probably the most valuable and demanded resource. We live in an internet worked society that relies on the distribution and circulation of information in the private as well as in the public sectors. Governmental, public, and private institutions are increasingly required to make their data electronically available. So here need to protect the privacy of the respondents (personals, organizations, associations, business establishments, and so on) [8].

A data stream is a sequence of unbounded, real time data items with a very high data rate that can only read once by an application. Assume a satellite remote sensor that is continuously generating data. The data are in very large amount (e.g. terabytes in volume), temporally ordered, quick changing, and potentially infinite. These benefits cause challenging issues in data streams field. We can say that Data Stream mining refers to informational structure extraction as models and patterns from continuous data streams. Data Streams mining have different issues in many aspects, such as computational, storage, querying and mining [1].

2. LITERATURE SURVEY

The literature survey is done to get detail knowledge of the of privacy preserving data mining. It is necessary to identity the various approaches and techniques that could be possibly used to preserve the sensitive data and private data. The objective of literature survey is to identity existing privacy preserving techniques, its advantages and dis-advantages and to find best suitable approach for preserving sensitive or private data [2].

2.1 Anonymization Based PPDM

The basic main form of the data in a table consists of four types of attributes:

(i) Explicit Identifiers is a set of attributes have information that identifies a record owner clearly such as name, SS number etc.

(ii) Quasi Identifiers is a set of attributes that could potentially identify a record owner when linked with publicly available data. Such as DOB, Sex, Zip etc.

(iii) Sensitive Attributes is a set of attributes that contains sensitive personal information such as salary, disease etc.

(iv) Non-Sensitive Attributes is a set of attributes that contains information that creates no problem if revealed even to unknown parties [4].

Anonymization refers to an approach where identity or/and sensitive data or private data about record owners are to be concealed. It even assumes that sensitive data or private data should be retained for analysis. It's understand that explicit identifiers should be removed but still there is a chance of danger of privacy intrusion when quasi identifiers are combined to publicly available data. This type of attacks are called as linking attacks. For example attributes such as Zip, Race, Sex, and DOB are available in public records such as voter list.

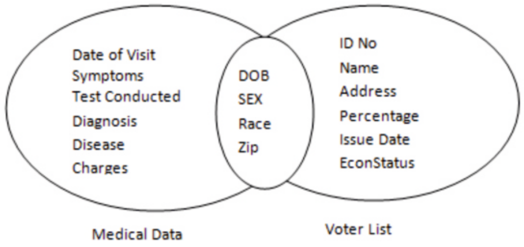

Figure:1 Linking Attack

Such records are also available in medical records, when combined, can be used to conclude the identity of the corresponding individual with high possibility as shown in figure:1.

Sensitive data or private data in medical record is disease or even medication prescribed. The quasi-identifiers have data like Zip, Race, Sex, DOB etc. are available in medical records and also this data in voter list that is publicly available. The explicit identifiers have data like Name, SS number etc. have been eliminated from the medical records.

Still, identity of individual can be predicted with higher possibility. Sweeney [9] proposed k-anonymity model which have two different methods like generalization and suppression to perform k-anonymity i.e. here any single record is distinguishable from at least k-1 other ones with respect to quasi-identifier attribute available in the anonymized dataset. In other words, we can outline a table as k-anonymous if the P1 values of each raw are equivalent to those of at least k- 1 other rows. Replacing a value with less specific but semantically consistent value is called as generalization and suppression involves blocking the values. Releasing this type data for mining reduces the risk of identification when combined with publically available data. But, at the same time, accuracy of the applications on the transformed data is reduced. A number of algorithms and different techniques have been proposed to implement k-anonymity using generalization and suppression in recent times.

Although the anonymization method makes sure that the transformed data is true but the problem is that heavy information loss. Moreover it is not resistant to homogeneity attack and background knowledge attack practically [10]. The first drawback of the k-anonymity model stem from the two conventions. First, it may be very hard for the owner of a database to decide which of the attributes are available or which are not available in outer tables. The second drawback is that the k-anonymity model acquires a certain method of attack, while in real situations; there is no reason why the intruder should not try other methods. However, as a research direction, k-anonymity is mixed-up with other privacy preserving methods needs to be analyzed, for detecting and even blocking k-anonymity violations.

2.2 Perturbation Based PPDM

Perturbation being used in statistical revelation control as it has an intrinsic property of simplicity, efficiency and ability to reserve statistical information. In this method the original values are changed with some artificial data values so that the statistical information computed from the perturbed data does not differ from the statistical information computed from the original data to a larger extent. The perturbed data records do not match to real-world record holders, so the attacker cannot perform the thoughtful linkages or recover sensitive knowledge from the available dataset. Perturbation can be performing by using adding noise or data interchanging or artificial data generation.

In the perturbation approach any distribution based data mining algorithm works under an indirect supposition to treat each dimension independently. Related information for data mining algorithms such as classification remains hidden in inter-attribute correlations. This is because the perturbation method acts different attributes independently. Hence the distribution based data mining algorithms have an inherent drawback of loss of hidden information available in multidimensional records. Another part of privacy preserving data mining that handles the drawback of perturbation method is cryptographic methods.

2.3 Randomized Response Based PPDM

Basically, randomized response based method is statistical technique introduced by Warner to solve a survey problem. In Randomized response based method, the data is perverted in such a way that the third party cannot say with chances better than a predefined threshold, whether the data from a customer contains right information or wrong information. The information received by each individual user is perverted and if the number of users is large, the collection information of these users can be determined with good quantity of accuracy. This is very important for decision-tree classification. It is based on mixed-up values of a dataset, somewhat individual data items. The data collection process in randomization method is carried out using two steps [9]. During first step, the data providers randomize their data and transfer this data to the data receiver. During next step, the data receiver rebuilds the original distribution of the data by using a distribution reconstruction algorithm. The randomization response model is shown in figure. 2.

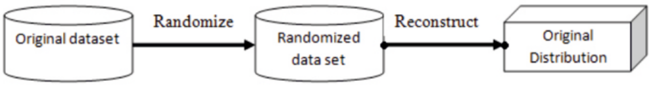

Figure: 2 Randomization Response Model

Randomization response method is very simple and does not need any knowledge of the distribution of other records in the data. Hence, the randomization response method can be executed at data collection time. It does not need a trusted server to contain the whole original data records in order to perform the anonymization process [11]. The drawback of a randomization method is that it acts all the data records equal irrespective of their local density. These indicate to an issue where the outlier records become more subject to oppositional attacks as compared to records in more compressed regions in the data [6]. One main point to this is to be having no purpose of adding noise to all the data records in the dataset. That reason it reduces the utility of the data for mining as the reconstructed distribution may not yield results in conformity of the purpose of data mining.

2.4 Condensation Approach Based PPDM

Condensation method builds constrained clusters in dataset and then produces pseudo data from the statistics of these clusters. It is called as condensation method because of its approach of using compressed statistics of the clusters to produce pseudo data. It creates sets of unlike size from the data, such that it is sure that each record lies in a set whose size is at least alike to its anonymity level. Here advanced, pseudo data are produced from each set so as to create a artificial data set with the same collection distribution as the unique data. Right now this method can be effectively used for the classification problem. Here use of pseudo-data provides an extra layer of protection, as it becomes hard to perform adversarial attacks on artificial data. Moreover, the aggregate behavior of the data is preserved, making it useful for a variety of data mining problems [11]. This approach is very useful in better privacy preservation as compared to other methods as it uses pseudo data rather than altered data. Moreover, it works even without changing data mining algorithms since the pseudo data has the same format as that of the original dataset. This method is very effective in case of data stream mining problems where the data is constantly changes. Here at the same time, data mining results get influenced as large amount of data is released because of the compression of a larger number of records into a single statistical group entity [7].

2.5 Cryptographic Based PPDM

Suppose a condition like that where multiple medical institutions wish to conduct a joint research for some mutual benefits without disclosing unnecessary data. In this condition, research regarding manifestation, discovery and medication based on various arguments is to be conducted and at the same time privacy of the individuals is need to be protected. Such conditions are referred to as distributed computing scenarios .The different parties involved in data mining of such tasks can be mutual untrusted parties, competitors; therefore protecting privacy becomes a major issue. We can say that Cryptographic methods are ideally meant for such environment where so many parties want integrate to compute results or share non sensitive data mining results and thereby avoiding disclosure of sensitive or private information. Cryptographic techniques find its utility in such scenarios because of two reasons: First, it offers a well-defined model for privacy that includes methods for proving and quantifying it. In second a huge set of cryptographic algorithms and constructs to execute privacy preserving data mining algorithms are available in this domain. The data may be sharing among different partners vertically or horizontally [3].

All these techniques are based on a special encryption protocol known as Secure Multiparty Computation (SMC) technology. SMC used in distributed privacy preserving data mining consists of a set of secure sub protocols that are used in horizontally and vertically partitioned data: secure sum, secure set union, secure size of intersection and scalar product. Although cryptographic methods ensure that the transformed data is complete and secure but this method fails to deliver when more than a few parties are involved in computation. Moreover, the data mining results may breach the privacy of individual records. There exist a so many number of solutions in case of semi-honest models but in the case of malicious models very less studies have been made [5][13].

3. PROPOSED APPROACH

Many techniques are present for privacy preserving in data mining but they have some shortcomings like information loss and data utility. This research work is mainly focus on using Perturbation techniques to preserve the privacy, increase data accuracy and decrease information loss.

First of all here data stream generated by MOA or take dataset from UCI data repository. The goal is to transform a given data set S into modified version S' that satisfies a given privacy requirement and preserves as much information as possible for the intended data analysis task. But in existing system make combination of multiple column value (Numeric value) then provide privacy to only one column value. So using new proposed method Sliding Window method we remove this drawback of existing system and provide privacy to individual column value using only one column.

Now apply classification method (Hoeffding Tree) for data stream mining on perturb dataset S'. So its generate classification model and then compare both result with respect to various evaluation parameters. In this way proposed method increase the accuracy of data stream classification and provide privacy to this data steam also.

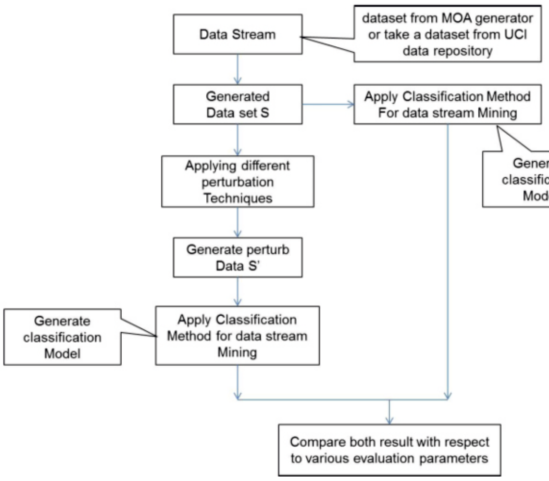
Data Stream
dataset from MOA generator or take a dataset from UCI data repository
Generated Data set S
Apply Classification Method For data stream Mining
Generate classification Model
Applying different perturbation Techniques
Generate perturb Data S'
Generate classification Model
Apply Classification Method for data stream Mining
Compare both result with respect to various evaluation parameters

Figure: 3. Framework for privacy preserving in data stream classification

See following table in that table there are 3 numeric attribute (Age, Salary, and Education Level) and 2 non-numeric attribute (Name and Gender)

TABLE I. ORIGINAL DATASET

Name	Age	Gender	Salary	Bonus
James	25	M	25000	4563.45
Bob	22	M	34000	2314.34
Alice	24	F	23400	3498.56
Prince	28	M	34500	4467.00

Numeric attribute is 3.Suppose selected attribute is salary.

First for first row:
- So add all this ex. 25 + 25000 + 4563.45 = 29588.45
- Mean= 29588.45/3= 9862.81
- So replace salary attribute values 25000 by mean value that is 9862.81

Then second row:
- So add all this ex. 22 + 34000 + 2314.34 = 36334.34
- Mean= 36334.34/3= 12112.11
- So replace salary attribute values 34000 by mean value that is 12112.11

After complete the calculation of all row output dataset will be like following:

TABLE II. OUTPUT OF EXISTING SYSTEM

Name	Age	Gender	Salary	Bonus
James	25	M	9862.81	4563.45
Bob	22	M	12112.11	2314.34
Alice	24	F	8974.18	3498.56
Prince	28	M	12998.33	4467.00

Now we provide privacy using Proposed approach. Numeric attribute is 3.Suppose selected attribute is salary.

First for first row:
- Window size is 3.
- So add all this ex. 25000 + 34000 + 23400 = 82400
- Mean= 82400/3= 27466.66
- So replace salary attribute values 25000 by mean value that is 27466.66

Then second row:
- So add all this ex. 34000 + 23400+34500 = 91900
- Mean= 91900/3= 30633.33
- So replace salary attribute values 34000 by mean value that is 30633.33

After complete the calculation of all row output dataset will be like following:

TABLE III. OUTPUT OF PROPOSED SYSTEM

Name	Age	Gender	Salary	Bonus
James	25	M	27466.66	4563.45
Bob	22	M	30633.33	2314.34
Alice	24	F	27633.33	3498.56
Prince	28	M	31166.66	4467.00

Now we have to perform classification on this perturb dataset using Hoeffding Tree and classified the data. We also make experiment on Bank Marketing Dataset - Bank marketing dataset taken from UCI dataset repository is related with direct marketing campaigns of a Portuguese banking institution, and it contain 45211 instances and 17 attributes.

4. CONCLUSION

The main purpose of privacy preserving data mining is developing algorithm to hide or provide privacy to certain sensitive or private information so that they cannot be disclosed or easily accessed to unauthorized parties or intruder. So based on experiment we can says that Existing system provide privacy to data stream but not provide accuracy to the dataset so this proposed approach remove the drawback of existing system and provide privacy to data stream and also increase the accuracy of the dataset .Here we provide privacy to only numeric value so in Future works we can extend this work and provide privacy to non-numeric value also.

ACKNOWLEDGMENT

I am very obliged to Dr. N. D. Shah, Principal of Parul Institute of Engineering and Technology for providing facilities to achieve the desire milestone. I also extend my thanks to Head of Department Dr. Gokulnath K for his inspiration and continuous support. I wish to warmly thank my guide, Ms. Tejal Patel for all her diligence, guidance, encouragement, inspiration and motivation throughout. Without her treasurable advice and assistance it would not have been possible for me to attain this landmark. I take this opportunity to thank all my friends for their support and help in each and every aspect.

REFERENCES

[1]. Kiran Patel, Hitesh Patel, Parin Patel, “Privacy Preserving in Data stream classification using different proposed Perturbation Methods ”, IJEDR, 2014, Volume 2, Issue 2 | ISSN: 2321-9939.

[2]. Manish Shannal, Atul Chaudhar, Manish Mathuria, Shalini Chaudhar, Santosh Kumar, “An Efficient Approach for Privacy Preserving in Data Mining ”, International Conference on Signal Propagation and Computer Technology, IEEE 2014,244-249.

[3]. Radhika Kotecha, Sanjay Garg, “Data Streams and Privacy: Two Emerging Issues in Data Classification”, 5th Nirma University International Conference on Engineering (NUiCONE), IEEE 2015.

[4]. Rupinder Kaur and Meenakshi Bansalt, “Transformation Approach for Boolean Attributes in Privacy Preserving Data Mining ” 1st International Conference on Next Generation Computing Technologies, IEEE 2015,644-648.

[5]. Neha Pathak, Shweta Pandey, “An Efficient Method for Privacy Preserving Data Mining in Secure Multiparty Computation”, Nirma University International Conference on Engineering, IEEE 2013,1-3.

[6]. Dhanalakshmi.M, Siva Sankari.E “Privacy Preserving Data Mining Techniques-Survey”, ICICES, IEEE 2014, ISBN No.978-1-4799-3834-6/14.

[7]. Hina Vaghashia, Amit Ganatra “A Survey: Privacy Preserving Techniques in Data Mining ”, International Journal of Computer Applications (0975 – 8887) Volume 119 – No.4, June 2015

[8]. C. Clifton, M. Kantarcioglu, and J. Vaidya, “Defining Privacy for Data Mining”, Next Generation Data Mining, AAAI/MIT Press, 2004.

[9]. Sweeney L, "Achieving k-Anonymity privacy protection uses generalization and suppression" International journal of Uncertainty, Fuzziness and Knowledge based systems, 10(5), 571-588, 2002.

[10]. Gayatri Nayak, Swagatika Devi, "A survey on Privacy Preserving Data Mining: Approaches and Techniques", ternational Journal of Engineering Science and Technology, Vol. 3 No. 3, 2127-2133, 2011.

[11]. Charu C. Aggarwal, Philip S. Yu “Privacy-Preserving Data Mining Models and algorithm” advances in database systems 2008 Springer Science, Business Media, LLC

[12]. http://www.tutorialspoint.com/data_mining/data_mining_tutorial.pdf

[13]. Jiawei Han, Micheline Kamber, Jian Pei. Data Mining Concepts and Techniques: 3rd Edn; Morgan Kaufmann Publishers is an imprint of Elsevier. 225 Wyman Street, Waltham, MA 02451, USA.